Preprints 2021 der Zeitschrift *BIBLIOTHEK – Forschung und Praxis*, BFP.2020.0115 **Hobert, Haupka und Jahn**

Anne Hobert, Nick Haupka und Najko Jahn

Entwicklung und Typologie des Datendiensts Unpaywall

Zusammenfassung: Analysen im Bereich des Open Access-Publizierens haben sich mit der Verfügbarkeit großer vernetzter Datensammlungen wie Unpaywall bedeutend vereinfacht. Der Artikel untersucht die Entwicklung des Datenbestands und der -struktur seit 2018. Eine Vollerhebung der Zeitschriftenartikel des Zeitraums 2008-18 zeigt, dass der OA-Anteil kontinuierlich wächst. Allerdings variiert die OA-Kategorisierung, was methodische Fragen beim Publikationsmonitoring und in der bibliometrischen Forschung aufwirft.

Schlüsselwörter: Open Access, Unpaywall, Publikationsmonitoring, Datenanalyse, Big Scholarly Data

Development and Typology of the Data Service Unpaywall

Abstract: Analyses in the field of open access publishing have become significantly easier with the availability of large connected data collections such as Unpaywall. The article examines the development of the database and its structure since 2018. A full survey of journal articles from the period 2008-18 shows that the OA share is growing continuously. However, the OA categorization varies, which raises methodological questions in publication monitoring and bibliometric research.

Keywords: Open access, Unpaywall, open access monitoring, data analysis, big scholarly data

1 Einleitung

Die bibliothekarische Diskussion über die Umstellung wissenschaftlicher Journale in den Open Access (OA) ist eng verbunden mit der Frage nach dem Anteil frei verfügbarer Artikel am Publikationsaufkommen. Während Bibliotheken zu Beginn der Entwicklung von OA-Dienstleistungen bibliometrische Kennzahlen für ihre Einrichtung oder ein Verlagsportfolio nur sehr aufwendig erheben konnten, hat in den letzten Jahren die Verfügbarkeit großer vernetzter Datensammlungen über wissenschaftliche Veröffentlichungen ein datengestütztes Berichtswesen befördert. Besonders wirkmächtig ist hierbei der OA-Discovery-Service Unpaywall[1], der mittlerweile in bibliometrischen Datenbanken wie dem Web of Science oder Scopus und in bibliothekarischen Discovery-Lösungen

[1] https://unpaywall.org/

integriert ist. Zudem fungiert Unpaywall als Datengrundlage einschlägiger OA-Studien aus Forschung und bibliothekarischen Praxis.

Der Beitrag beinhaltet eine Datenanalyse des Services Unpaywall. Vor dem Hintergrund der bibliometrischen Informationsbedürfnisse wissenschaftlicher Bibliotheken in Deutschland im Kontext der OA-Transformation, beschreiben wir Unpaywalls Datenstruktur und untersuchen, wie sich der Datenservice seit 2018 entwickelt hat. Insbesondere fragen wir, nach welchen Kriterien und mittels welcher Methoden OA-Volltexte identifiziert wurden. Auf Grundlage von zehn Datendumps, die Unpaywall seit 2018 offen verfügbar gemacht hat, lassen sich nicht nur allgemeine Entwicklungen beim OA-Publizieren nachzeichnen, sondern auch begriffliche und methodologische Herausforderungen bei der Analyse des OA-Publikationsaufkommens unter Verwendung von Unpaywall illustrieren.

2 Publikationsmonitoring an wissenschaftlichen Bibliotheken

Mit der Einführung von Services zur Unterstützung des OA-Publizierens fanden auch datenanalytische Verfahren unter dem Stichwort „Publikationsmonitoring" Eingang in das Tätigkeitsspektrum wissenschaftlicher Bibliotheken in Deutschland. Laut Schmeja und Tullney (2020) dient das Publikationsmonitoring der datengestützten Entscheidung über Entwicklungen im Bereich des OA-Publizierens mit Schwerpunkt auf die Anzahl und den Anteil frei verfügbarer Zeitschriftenartikel einer wissenschaftlichen Einrichtung oder eines Forschungsverbunds. Das Publikationsmonitoring findet zudem Anwendung in wissenschafts- und förderpolitischen Beratungskontexten, etwa bei der Anbahnung und Evaluierung von OA-Mandaten.[2]

Ein wesentlicher Treiber des Publikationsmonitorings durch wissenschaftliche Bibliotheken in Deutschland war das Förderprogramm „Open Access Publizieren" der Deutschen Forschungsgemeinschaft (DFG), welches seit 2011 die Etablierung und Weiterentwicklung von universitären OA-Publikationsfonds unterstützt. Dabei lag das Augenmerk der DFG bis Ende 2020 auf der anteiligen Förderung von Publikationsgebühren für Artikel in reinen OA-Zeitschriften. Hochschulen waren als Antragstellerinnen nicht nur dazu aufgefordert, die OA-Artikel ihrer Einrichtung zu erheben und ins Verhältnis zum institutionellen Gesamtpublikationsaufkommen zu setzen. Sie sollten zudem über die Methoden und zugrunde liegenden Datenquellen Auskunft geben,

[2] Ein prominentes Beispiel sind die OA-Zielvorgaben für den Europäischen Forschungsraum von 2012 und die darauffolgende Implementierung des OA-Mandats für Drittmittelprojekte des Forschungsrahmenprogramms HORIZON 2020.

um eine fachliche Diskussion über die Erhebung von Daten über OA-Veröffentlichungen zu ermöglichen.[3]

Laut einer Evaluierung des DFG-Programms zeichnet sich das OA-Berichtswesen durch eine hohe Heterogenität aus, während die Anforderungen im Zuge von OA-Transformationsverträgen gestiegen sind.[4] Die einschlägige Initiative *ESAC*[5] definiert OA-Transformationsverträge als Lizenzmodelle, bei denen der lesende Zugang und die Publikation in einer Zeitschrift gemeinsam betrachtet werden. Ziel ist eine Umstellung des Geschäftsmodells wissenschaftlicher Verlage und damit auch der bibliothekarischen Erwerbung von Subskriptionen („Bezahlzugang") zur finanziellen Förderung der OA-Publikation.[6] Die wirkmächtige Transformationsinititative *OA 2020* betont dabei die Notwendigkeit einer belastbaren Datengrundlage, um OA-Transformationsverträge weiterzuentwickeln und zu evaluieren.[7] Entsprechend beabsichtigt die internationale Förderinitiative Plan S, die ihr Vorgehen in Übereinstimmung mit der OA2020-Initiative sieht, ein Monitoring, ob und inwieweit OA-Fördermandate umgesetzt werden.[8]

Bibliotheken und ihre Konsortien können mittlerweile auf eine Vielzahl an Datenquellen und -services für das Publikationsmonitoring zurückgreifen. Besonders bemerkenswert ist die Rolle von *Big Scholarly Data*, definiert als schnell wachsende große Datensätze über wissenschaftliche Veröffentlichungen[9], beim Nachweis von OA-Volltexten. Die Bibliometrie unterscheidet Anbieter für Big Scholarly Data bezüglich der Frage, welche Publikationsorte ausgewertet werden und hinsichtlich des Geschäftsmodells einschließlich der Zugangsbedingungen und Möglichkeiten der Nachnutzung der zugrundeliegenden Datenbasis.[10] Kommerzielle bibliometrische Datenbanken wie das Web of Science, Scopus oder Dimensions, die ihre Auswertungsmöglichkeiten um Evidenzen für offen verfügbare Volltexte erweitert haben, werden mittlerweile durch frei verfügbare Angebote wie Microsoft Academics, Semantic Scholar, PubMed oder dem OpenAIRE Research Graph ergänzt, deren Datenbestand frei und maschinenlesbar verfügbar ist. Auf der Grundlage von *Big Scholarly Data* entstanden zudem neuartige OA-Monitoringangebote für wissenschaftliche Einrichtungen wie das Leiden Ranking[11] oder deutsche Open Access Monitor[12]. Ebenfalls ist zu beobachten, dass

[3] Siehe Fournier und Weihberg (2013) für eine ausführliche Darstellung des Programms.
[4] Barbers et al. (2020).
[5] https://esac-initiative.org/
[6] https://esac-initiative.org/about/transformative-agreements/
[7] https://oa2020.org/ ; siehe auch das zugrunde liegende White Paper der Initiative von Schimmer et al. (2015).
[8] https://www.coalition-s.org/guidance-on-the-implementation-of-plan-s/
[9] Xia et al. (2017).
[10] Waltman und Larivière (2020).
[11] Robinson-Garcia (2020).

einzelne Verlage wie Elsevier verstärkt Data-Analytics-Services mit OA-Transformationsverträgen verzahnen.[13]

Aufgrund der vielfältigen Datenakteure haben Bibliotheken und ihre Konsortien Beratungsbedarf bei der Einbindung neuartiger Services in die Planungs- und Geschäftsprozesse.[14] Konkret bezieht er sich auf die Kategorisierung der OA-Nachweisinformation, den Umfang und der Nachvollziehbarkeit der Datenbasis sowie Tools und benötigte Kompetenzen zur Datenauswertung.[15] Aufgrund seiner breiten Nutzung beim Publikationsmonitoring steht insbesondere der OA-Nachweisservice Unpaywall des Non-Profit Unternehmen *Our Research*[16] (vormals *Impactstory*) im Fokus des Interesses. Unpaywall bietet Services zum automatisierten Auffinden einzelner OA-Volltexte. Aufgrund der freien Verfügbarkeit der Daten ist Unpaywall in Forschung und Praxis für Datenanalysen weit verbreitet. Beispiele im deutschen Bibliothekswesen umfassen den OA-Monitor des FZ Jülich[17] und die jährlichen OA-Monitoringberichte für das Land Berlin[18]. Die bibliometrischen Datenbanken Web of Science, Scopus und Dimensions haben Unpaywall. Zugleich wird Unpaywall auch für quantitative OA-Studien mit besonderem Fokus auf das institutionelle Publikationsverhalten verwendet. Während in der Literatur insbesondere die Selektivität des Services hinsichtlich der ausgewerteten OA-Quellen und die Typologisierung der OA-Volltextnachweise andiskutiert wird[19], fehlt es an Untersuchungen, die die Entwicklung der OA-Nachweiskategorien und ihres Vorkommens in Unpaywall nachzeichnen. Im Folgenden werden wir daher mittels einer Datenanalyse die Entwicklung Unpaywalls seit 2018 auf Basis von zehn Datensnapshots nachzeichnen.

3 Herangehensweise

Unpaywall basiert auf Crossref-indexierten Publikationen. Crossref ist eine DOI-Registrierungsagentur, über die Verlage Metadaten zu ihren registrierten Publikationen hinterlegen. Datenqualität und -umfang hängen dabei stark davon ab, welche Informationen von einzelnen Verlagen bereitgestellt werden und wie sorgfältig bei der Eintragung von Metadaten vorgegangen

[12] Mittermaier (2018).
[13] Aspesi und Brand (2020).
[14] Hillenkötter (2018).
[15] Schmeja und Tullney (2020).
[16] https://ourresearch.org/
[17] MIttermaier et al. (2018).
[18] Kindling et al. (2020).
[19] Siehe die Diskussion über die Selektivität und Repositorienkategorisierung Unpaywalls bei Robinson-Garcia et al. (2020) und Huang et al (2020).

wird.[20] Unpaywall ermittelt für alle bei Crossref registrierten Publikationen frei verfügbare Volltexte und erschließt die OA-Informationen.[21] Die OA-Bestandsnachweise werden dann in Form verschiedener Serviceangebote zur Verfügung gestellt, die auf unterschiedliche Anwendungsszenarien zugeschnitten sind. So ermöglichen eine REST API und eine webbasierte Suchmaske das Abrufen von OA-Nachweisen für eine begrenzte Anzahl an DOIs. Für den Zugriff auf die API stehen Softwarepakete für die im Bereich Datenanalysen populären Programmiersprachen R[22] und Python[23] zur Verfügung, die wir an der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB Göttingen) mitentwickeln. Zudem gibt es eine Browsererweiterung, die Links zu frei verfügbaren Volltextversionen von kostenpflichtigen Artikeln ermittelt. Das Abonnement des kostenpflichtigen Data Feeds beinhaltet wöchentlich aktuelle Updates des gesamten Datensatzes. Für unsere Zwecke - die Durchführung bibliometrischer Studien zur freien Verfügbarkeit von wissenschaftlichen Publikationen - eignen sich am besten die Datensnapshots, die mehrmals im Jahr von Unpaywall veröffentlicht werden und kostenlos zum Download zur Verfügung stehen. Sie enthalten statische Momentaufnahmen des vollständigen Unpaywalldatensatzes, wodurch eine Vergleichbarkeit zwischen Analysen ermöglicht wird.

Im Folgenden werten wir die statischen Datensnapshots von Unpaywall aus und stellen dafür unseren Datenverarbeitungsprozess vor (siehe Abbildung 1). Ziel des Prozesses ist es, eine konsolidierte und hochperformante Datenbasis für OA-Studien an der SUB Göttingen bereitzustellen. Zu diesem Zweck werden die Datensätze, die etwa vierteljährlich als komprimierte JSON Newline Dateien mit einer Größe von ca. 20 GB (über 100 GB unkomprimiert) veröffentlicht werden und jeweils Informationen über den OA-Status von etwa 100 Millionen Publikationen enthalten, in einem ersten Schritt heruntergeladen. Um die Datensätze trotz ihres bedeutenden Umfangs effizient verarbeiten zu können, extrahieren wir anschließend die für unsere Anwendungszwecke relevanten Informationen mit Hilfe des command line tools jq[24]. Dabei filtern wir die enthaltenen Publikationen nach Publikationsdatum (enthalten sind die Jahre ab 2008) und entfernen einige umfangreiche Felder, die für die Analysen nicht benötigt werden. Die so reduzierten Datensätze laden wir anschließend in die hochperformante Analyseumgebung von Google BigQuery[25], was uns eine zeiteffiziente Abfrage und Manipulation der Daten mittels der

[20] Hendricks et al. (2020).
[21] Piwowar et al. (2018).
[22] Jahn (2019).
[23] Haupka und Morrison (2020).
[24] https://stedolan.github.io/jq/
[25] https://cloud.google.com/bigquery

Abfragesprache SQL ermöglicht. Das Cloud-basierte Angebot erleichtert darüber hinaus die Zusammenarbeit im Team, weil die Authentifizierung und das Rechtemanagement Google-Konten erfolgt. BigQuery können wir über bereits vorhandene Schnittstellen und Softwarepakete in unsere gewohnten Analyseprozesse, basierend auf R bzw. Python, integrieren. Die monatlichen Kosten für den Cloud-Dienst sind abhängig von der Nutzung und betragen bei uns im Schnitt ein bis zwei Euro monatlich.

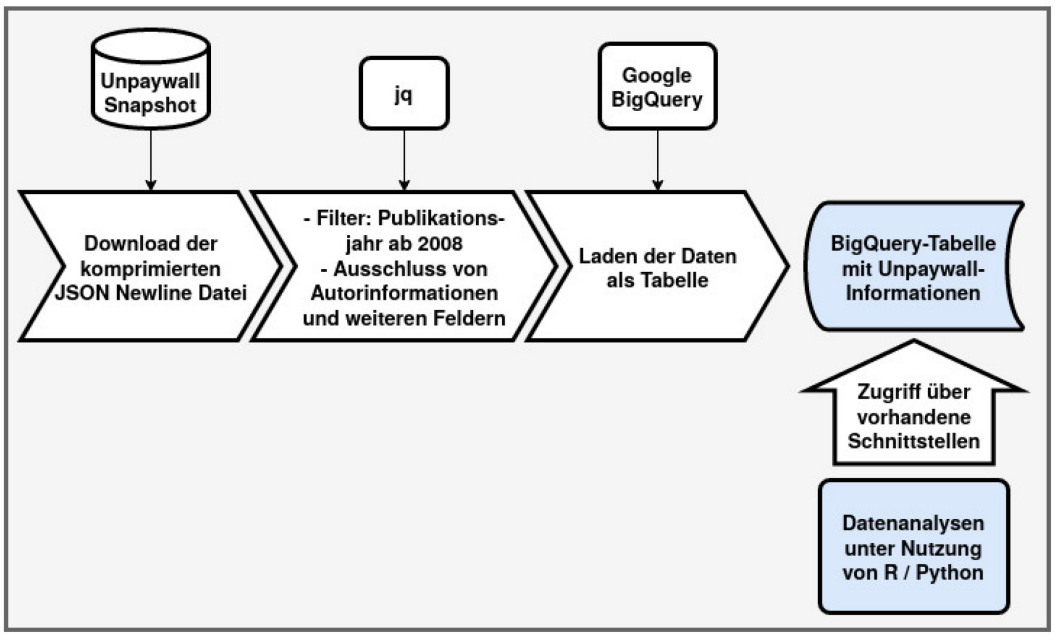

Abb. 1: Schematische Darstellung der Vorgehensweise zur Nutzung der frei zugänglichen Unpaywall-Snapshots für Datenanalysen.

4 Ergebnisse

Nachfolgend präsentieren wir auf Grundlage der oben skizzierten cloud-basierten Dateninfrastruktur die Entwicklung des Datenbestands Unpaywalls. Bei unserer Analyse betrachten wir ausschließlich Zeitschriftenartikel, die in den Jahren zwischen 2008 und 2018 veröffentlicht wurden. Darauf aufbauend erfolgt eine Übersicht der bisher erschienen Datensnapshots sowie eine Auswertung der wesentlichen Merkmale und Änderungen zwischen den Dumps. Anschließend sollen auf Basis des Snapshots von April 2020, Aussagen über das OA-Publikationsverhalten getroffen werden. Insbesondere wird der Aspekt der Mehrfacharchivierung genauer betrachtet, also ob und inwiefern OA-Volltexte über mehrere Publikationsorte verfügbar sind. Unpaywall eignet sich für solche Analysen, da der Datendienst für jeden indexierten Artikel sämtliche OA-Volltexte erfasst. Unpaywall hierachisiert die OA-Nachweise und ermittelt eine sogenannte *is_best* Publikation. Diese wird durch

einen speziellen Algorithmus selektiert, welcher unter anderem Verlagsangebote vor dem Versionszustand (z.B. Preprint, Postprint, finale Version) sowie vor der Archivierung auf einem Repositorium priorisiert.[26] Aufgrund der vielfätligen Quellen, die Unpaywall auswertet, spielt die Analyse der verwendeten OA-Kategorisierung eine nicht zu unterschätzende Rolle. Aus diesem Grund werden wir die Ergebnisse der OA-Typologisierung des diesjährigen Snapshots von 2020 mit denen eines Snapshots des Vorjahres 2019 kontrastieren. Abschließend erfolgt eine Untersuchung der zeitlichen Entwicklung der Evidenztypen. Sie zeigen an, auf welche Art und Weise Unpaywall OA-Volltexte identifiziert hat.

4.1 Zusammenfassung bisher erschienener Snapshots

Seit 2018 erstellt Unpaywall in regelmäßigen Abständen Datenbankdumps und macht sie öffentlich zugänglich. Bislang liegen zehn Snapshots vor, die jeweils eine unterschiedliche Repräsentation der Unpaywall-Datenbank zu festen Zeitpunkten dokumentieren. Der erste verfügbare Snapshot wurde im März 2018 bereitgestellt und enthält, neben Publikationen aus anderen Jahren, Metadaten von über 27,5 Millionen Zeitschriftenartikeln, die zwischen 2008 und 2018 mit einer Crossref-DOI publiziert wurden. Die Anzahl der indexierten Publikationen ist seitdem kontinuierlich gestiegen (siehe Tabelle 1). Der Snapshot von April 2020 umfasst Informationen über fast 31,5 Millionen Zeitschriftenartikel, was einer Steigerung des Datenbestands von mehr als 14% entspricht. Neben einem Anstieg an Artikeln, der über mehrere Snapshots zu beobachten ist, lassen sich auch diverse andere Änderungen am Datenbestand und dessen Struktur ausmachen. So sind Veränderungen am zugrunde liegenden Datenschema erkennbar, was zu einer Inkonsistenz zwischen verschiedenen Snapshots führt. Speziell betrifft das das Feld *x_reported_noncompliant_copies*, welches seit dem Juni 2018 Snapshot nicht mehr existiert und die Felder *oa_status*, *has_repository_copy*, *endpoint_id*, *journal_issn_l*, *repository_institution*, *is_paratext*, welche zu unterschiedlichen Zeitpunkten in die Unpaywall-Datenbank integriert worden sind. Das aktuelle Datenschema kann auf der Webseite von Unpaywall eingesehen werden[27].

Über das Feld *is_oa*, welches den OA-Status einer Publikation wiedergibt, kann der OA-Anteil eines Snapshots ermittelt werden. Der Anteil an frei verfügbaren Artikeln liegt demnach im März 2018 bei 32%, wohingegen der aktuelle Snapshot 43% verzeichnet (vgl. Tabelle 1).

Die OA-Kategorisierung einer Publikation wird von Unpaywall auf Basis verschiedener Verfahren, sogenannter Evidenztypen, ermittelt. Hier sind zum Teil leichte Schwankungen erkennbar. So

[26] Das Verfahren ist in Piwowar et al. (2018) ausführlich beschrieben.
[27] https://unpaywall.org/data-format

werden bei Zeitschriftenartikeln anfangs mehr als 20 unterschiedliche Methoden zur OA-Erschließung genutzt, während aktuellere Snapshots etwa 15 bis 20 Evidenztypen beinhalten. Auch das Verhältnis von Veröffentlichungen über Verlage und Repositorien, welches über das Feld *host_type* identifiziert werden kann, verändert sich über mehrere Snapshots hinweg. Hier kann ein leichter Anstieg von OA-Artikeln auf Repositorien beobachtet werden.

Tab. 1: Kurzübersicht über die wesentlichen Änderungen zwischen verschiedenen Unpaywall-Snapshots. Die Spalten beinhalten für jeden Snapshot Angaben zu den indexierten Publikationen (hier begrenzt auf Zeitschriftenartikel), Informationen zum Open Access Anteil, Angaben zum Verhältnis von Veröffentlichungen über Verlage und Repositorien (hierbei wurden nur Publikationen mit dem Attribut is_best=True berücksichtigt), Neuerungen im Datenschema sowie eine statistische Auswertung der Evidenztypen.

	Anzahl der enthaltenen Publikationen	Open Access Anteil	Repositorien / Verlage Anteil	Änderungen Datenschema	Anzahl der Evidenztypen
März 2018 Snapshot	27.498.666	31,56%	21,45% / 78,55%		22
April 2018 Snapshot	27.268.179	32,58%	19,80% / 80,20%		22
Juni 2018 Snapshot	28.201.468	35,55%	19,17% / 80,83%	x_reported_non compliant_copies entfällt	23
September 2018 Snapshot	29.908.032	37,86%	17,21% / 82,79%		15
Februar 2019	31.159.960	37,34%	17,97% / 82,03%	endpoint_id hinzugefügt	15

Snapshot					
April 2019 Snapshot	31.341.794	38,11%	17,68% / 82,32%	oa_status hinzugefügt	17
August 2019 Snapshot	31.602.128	38,75%	17,23% / 82,77%	journal_issn_l und repository_institution hinzugefügt	19
November 2019 Snapshot	31.827.129	42,63%	26,52% / 73,48%	has_repository_copy hinzugefügt	18
Februar 2020 Snapshot	31.611.299	42,79%	22,76% / 77,24%	is_paratext hinzugefügt	17
April 2020 Snapshot	31.482.103	42,50%	23,22% / 76,78%		18

4.2 Überschneidungsbereich Verlage/Repositorien

Unpaywall verzeichnet je Artikel sämtliche OA-Volltexte, die der Service identifizieren konnte. Das Listenelement *oa_location* enthält detaillierte Informationen zur OA-Quelle einschließlich einer Typologisierung. Hierbei unterscheidet Unpaywall unter anderem zwischen Verlagsangeboten und Repositorien. Daraus lassen sich die Anteile der jeweiligen Hosting Locations berechnen. Im aktuellen Snapshot vom April 2020 sind demnach 77% aller Zeitschriftenartikel über die Website eines Verlages erhältlich und etwa 62% der in Unpaywall indexierten Zeitschriftenartikel sind in Repositorien verfügbar. Dabei sind 38% der Publikationen sowohl auf einer Verlagswebsite als auch auf einem Repositorium abrufbar. Der Trend, welcher sich in Abbildung 2 abzeichnet, offenbart, dass

in den vergangenen Jahren die zusätzliche Archivierung eines Artikels auf einem Repositorium beliebter wird.

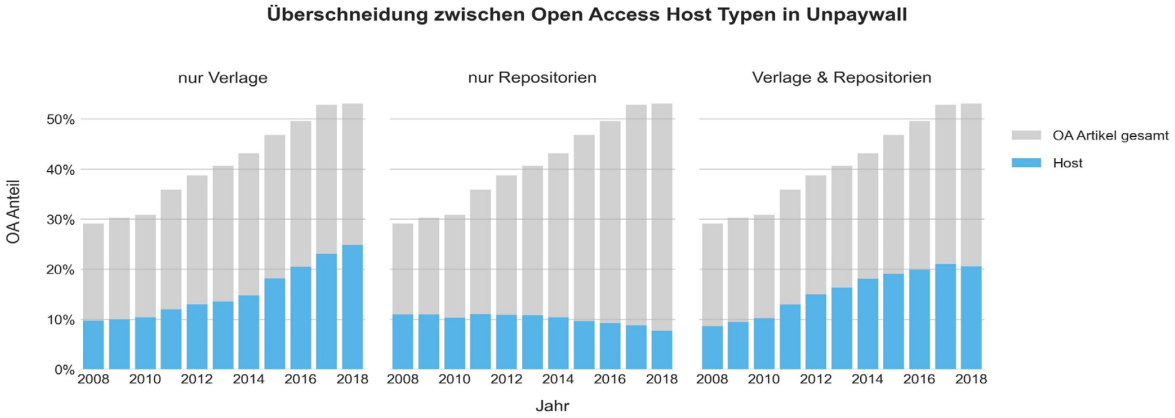

Abb. 2: Open Access Zeitschriftenartikel nach OA-Host. Die blauen Säulen stellen die Anzahl der Artikel pro Host dar, während die grauen Säulen die gesamte Anzahl der OA-Artikel Host-übergreifend darstellen. Datengrundlage sind Zeitschriftenartikel von 2008 bis 2018 im der Unpaywall-Snapshot von April 2020.

4.3 Kombinationen von Evidenztypen

Unpaywall kategorisiert OA-Nachweise zudem über sogenannte Evidenztypen. Weil die Kategorisierung je OA-Volltextlink erfolgt, ist eine Kombinationen verschiedener Evidenztypen möglich (siehe Abbildung 3). Beim Betrachten der häufigsten Kombinationen von Evidenztypen im Snapshot von April 2020 fällt auf, dass besonders häufig Überschneidungen zwischen Nachweisen aus verschiedenen Repositorien existieren. Tatsächlich ist der OA-Nachweis über OAI-PMH und PMC die vierthäufigste Kombinationsmöglichkeit im aktuellen Snapshot. Die drei am häufigsten vorkommenden Kombinationen sind einzelne Evidenztypen, wobei überwiegend Artikel über ein frei erhältliches PDF identifiziert werden, ohne dass jedoch eine freie Lizenz oder Informationen über Nutzungsrechte festgestellt werden können.

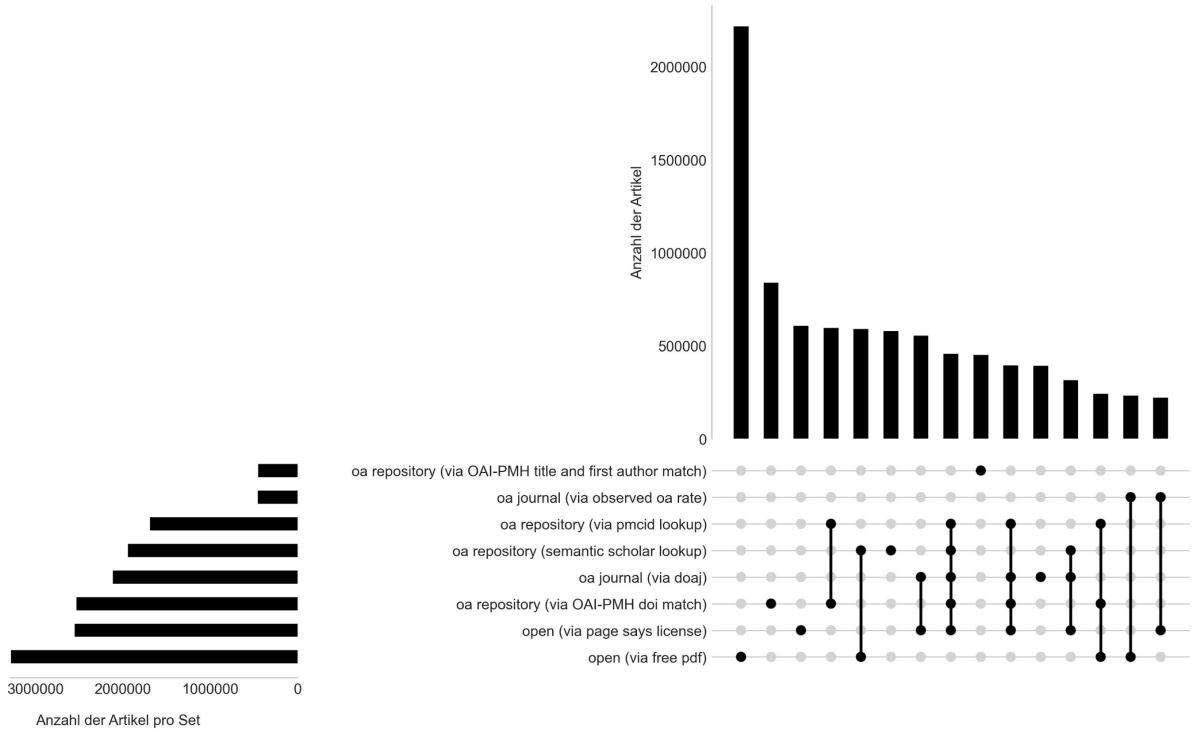

Abb. 3: Die fünfzehn häufigsten Kombinationen von Evidence-Typen im April 2020 Snapshot für Zeitschriftenartikel zwischen 2008 und 2018. Das Balkendiagramm auf der linken Seite stellt die Anzahl der Artikel pro Evidence-Typ dar ("Anzahl der Artikel pro Set"). Das Säulendiagramm auf der rechten Seite bildet die Anzahl der Artikel für die jeweiligen Kombinationsmöglichkeiten ab ("Anzahl der Artikel"). Die Überschneidung von Evidence-Typen wird durch schwarze Punkte im unteren Bereich der Grafik repräsentiert.

4.4 Verbreitung von Open Access Modellen

Unpaywall verwendet eine eigene OA-Klassifizierung mit der der OA-Status einer Publikation ermittelt werden kann.[28] Die OA-Klassifizierung von Unpaywall basiert auf einem deterministischen Algorithmus, welcher den OA-Typ einer Publikation bestimmt, der über das Feld *oa_status* abgerufen werden kann.[29] Der Algorithmus hat sich seit der Einführung im Februar 2019 Snapshot teilweise modifiziert, sodass es zum Teil zu starken Unterschieden zwischen Anteilen diverser OA-Typen in verschiedenen Snapshots kommt. Verglichen mit dem April 2019 Snapshot hat sich beim Snapshot von April 2020 die Identifizierung von Hybrid-OA erheblich verändert. Das ist besonders am Publikationsjahr 2018 erkennbar, für welches im 2019er Snapshot über 50% der OA-Publikationen als Hybrid-OA gekennzeichnet werden, wohingegen im 2020er Snapshot nur etwas

[28] Piwowar et al. (2018).

[29] https://support.unpaywall.org/support/solutions/articles/44001777288-what-do-the-types-of-oa-status-green-gold-hybrid-and-bronze-mean-

über 10% der OA-Zeitschriftenartikel als Hybrid-OA identifiziert werden. Umgekehrt hat sich die Klassifizierung von Gold-OA innerhalb der Snapshots geändert. So sind im früheren Snapshot im Jahre 2018 noch weniger als 10% der OA-Artikel Gold-OA, während im aktuelleren Snapshot mehr als 50% der OA-Publikationen als Gold-OA veröffentlicht wurden. Hintergrund ist eine Änderung bei der Identifizierung von reinen OA-Journalen. Wurde zunächst ausschließlich das DOAJ zur Identifizierung dieser Journale genutzt wurde, hat Unpaywall das Verfahren 2019 erweitert.[30] Dass Zeitschriften zu OA konvertieren ist vermutlich nachrangig, weil nur wenige Journale ihr Geschäftsmodell ändern.[31]

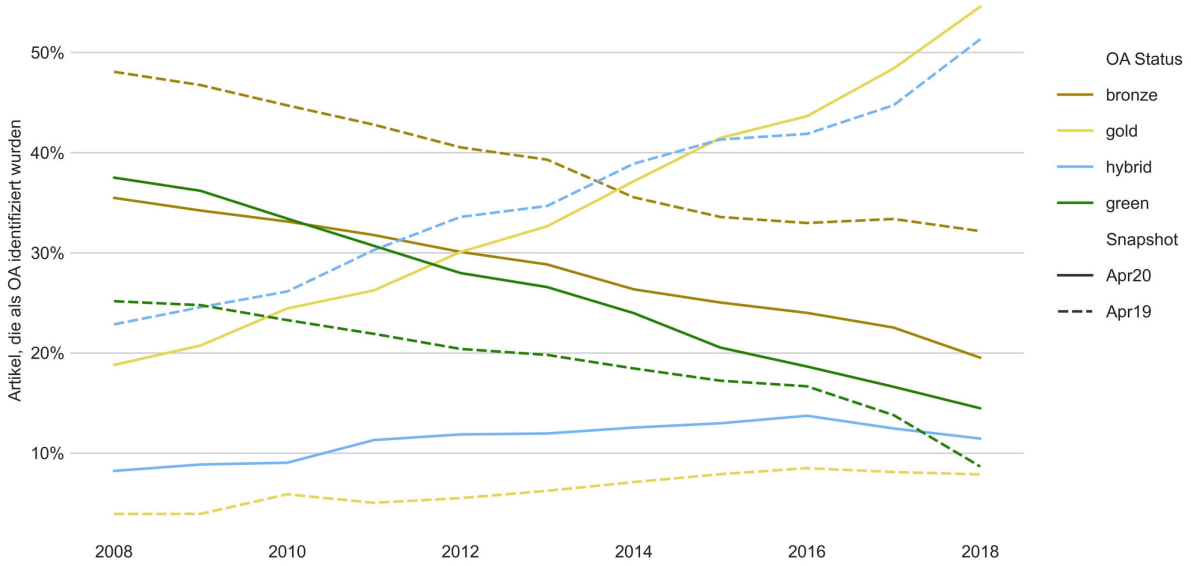

Abb. 4: Verbreitung von OA-Varianten nach Unpaywall. Die gestrichelten Linien zeigen den Verlauf auf Basis des Unpaywall Dumps von April 2019, wohingegen die durchgezogenen Linien die Daten aus dem April 2020 Snapshot repräsentieren. Betrachtet werden Zeitschriftenartikel zwischen 2008 und 2018, die laut Unpaywall als OA erhältlich sind. Closed-OA wurde im Sinne der Übersichtlichkeit nicht abgebildet.

4.5 Vorkommen von Evidenztypen

Das Feld *evidence* gibt Rückschlüsse, wie der OA-Status einer Publikation durch Unpaywall ermittelt worden ist. Zum Beispiel bedeutet *oa journal (via journal title in doaj)*, dass das Journal des gesuchten Artikels im Directory of Open Access Journals (DOAJ) registriert war. Die Werte der Variable sind laut *Our Research* fortlaufenden Änderungen unterworfen und nicht standardisiert, was wir nachfolgend konkretisieren: Im Laufe der letzten drei Jahre hat Unpaywall verschiedene OA-

[30] https://support.unpaywall.org/support/solutions/articles/44001792752-how-do-we-decide-if-a-given-journal-is-fully-oa-
[31] Momeni et al. (2019)

Identifikationsverfahren genutzt, wie in Abbildung 5 zu sehen ist. Hier wird erkenntlich, dass Unpaywall vor allem zu Beginn diverse Datenquellen wie etwa BASE und DataCite genutzt und unterstützt hat, diese jedoch nach geraumer Zeit aufgegeben hat. Neu hingegen ist der Evidenztyp für Hybrid-OA (*hybrid (via page says license)*), welcher im April 2019 Snapshot hinzugefügt worden ist. Auch die Ermittlung von OA via Semantic Scholar ((*oa repository semantic scholar lookup)*) ist eine Neuerung, die seit dem Snapshot von November 2019 existiert. Auffällig ist, dass der Evidenztyp *hybrid (via free pdf)* zwischenzeitlich nicht verwendet wurde, allerdings seit dem April 2020 Snapshot wieder auftaucht.

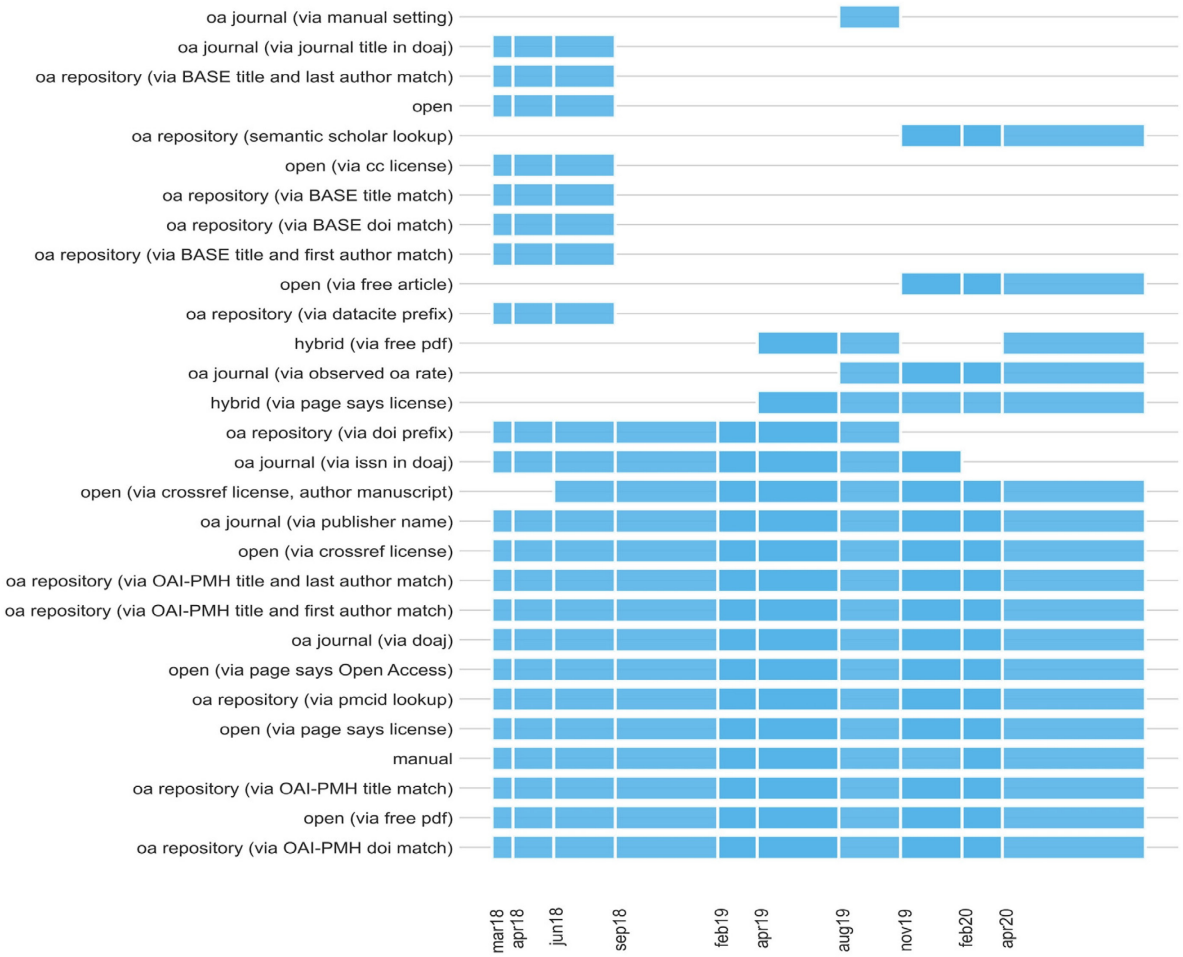

Abb. 5: Vorkommen von Evidenztypen in Unpaywall Snapshots. Die einzelnen Snapshots werden durch einen blauen Balken mit einer weißen Umrandung kenntlich gemacht. Die Evidenztypen sind nach der Häufigkeit des Auftauchens aufsteigend sortiert.

5 Diskussion

Insgesamt betreibt Our Research mit Unpaywall eine sehr nützliche Datenquelle über OA-Nachweise - sowohl für die praktische Anwendung als auch für wissenschaftliche Studien im Bereich der Bibliometrie. Bis auf den wöchentlich aktualisierten Data Feed sind die Angebote kostenfrei nutzbar. Die Entwicklung erfolgt als Open Source Software, was Transparenz schafft sowie die Möglichkeit, zur Weiterentwicklung oder Fehlerbehebung beizutragen. Zudem ist das eingesetzte Datenschema schlank und übersichtlich. Wichtige Informationen (wie etwa, ob für eine Publikation eine frei verfügbare Version gefunden wurde oder nicht) sind direkt aus den Daten abzulesen. Für detaillierte Analysen, wie eine eigenständige Klassifizierung des OA-Typs, stehen darüber hinaus grundlegende Informationen wie der Evidenztyp oder die URL der einzelnen Fundorte zur Verfügung, die sich für eigene Kategorien kombinieren lassen[32]. Wie an dem hier beispielhaft vorgestellten Workflow, die Unpaywall Datensätze mit Hilfe von Google BigQuery und freier Software wie R oder Python bereitzustellen und zu analysieren, ersichtlich wird, ist die Nutzung und Verarbeitung der Daten nur mit äußerst geringen Kosten verbunden. Diese auch bei intensiver Nutzung geringen monatlichen Kosten entstehen durch die Verwendung von BigQuery. Neben dem überschaubaren finanziellen Aufwand ist ein weiterer Vorteil dieser Herangehensweise, dass die Datenanalyse mit existierenden Data Science Tools möglich ist. Somit ist weder die Entwicklung eigener Software, noch ein eigenständiges Datenbankhosting erforderlich. Dies erleichtert zum einen die Arbeitsprozesse und setzt die Hemmschwelle zur Integration der Daten in eigene Arbeitsumgebungen herab. Zum anderen wird in Kombination mit der Offenlegung der verwendeten Softwarepakete und Datensnapshots auf diese Weise auch die Reproduzierbarkeit der erhaltenen Ergebnisse gewährleistet. Weiterhin erleichtert die Verwendung existierender Software und Cloud-basierter Angebote die Zusammenarbeit im Team.

Dennoch ergeben sich die folgenden Problembereiche in Bezug auf Datenanalysen mit Unpaywall. Zunächst stellt sich die Frage nach Nachhaltigkeit und Datenstabilität. Auch wenn bislang alle von Unpaywall veröffentlichten Snapshots kostenfrei zum Download verfügbar sind, ist deren Langzeitarchivierung nicht dauerhaft garantiert. Wir sichern daher die Datendumps, die wir für unsere Analysen verwenden, zusätzlich lokal an der SUB Göttingen. Weiterhin sind sowohl das Datenschema (siehe Tabelle 1) als auch die in den Daten enthaltenen Informationen selbst (siehe Abbildung 4) fortlaufend Veränderungen unterworfen[33]. Dies überträgt sich auf Ergebnisse und Schlussfolgerungen aus Datenanalysen basierend auf Unpaywalldaten, die in der Folge ebenfalls

[32] So wenden Hobert et al. (2020) ein eigen entwickeltes Kategorienschema auf Basis von Unpaywallmetadaten an.
[33] Vgl. Piwowar et al. (2019).

stark variieren können. Für eine Reproduzierbarkeit und Vergleichbarkeit von datenanalytischen Befunden ist daher die genaue Angabe des verwendeten Datensatzes sowie des Erhebungszeitpunktes entscheidend. Für Untersuchungen zu spezifischen OA-Typen kann außerdem die verwendete Typologisierung problematisch werden. Wie im Kapitel Ergebnisse beschrieben, bietet Unpaywall seit 2019 eine Kategorisierung von OA-Publikationen im Feld *oa_status* an. Diese berücksichtigt allerdings jeweils nur den als "beste Location" klassifizierten Fundort und priorisiert bei der Auswahl verlagsbasiertes OA (freie Verfügbarkeit über eine Host Location vom Typ publisher) immer gegenüber repositorienbasiertem OA (freie Verfügbarkeit über eine Host Location vom Typ repository). Der nicht unerhebliche Überschneidungsbereich (siehe Abbildung 2) wird dabei ausschließlich dem verlagsbasierten OA zugerechnet, was zu einer massiven Unterschätzung der Bedeutung von Repositorien für die freie Verfügbarkeit von wissenschaftlichen Publikationen führen kann. Zu bemerken ist allerdings, dass die Unpaywalldatensätze im *oa_locations* Objekt alle Fundorte auflisten und etwa über den Typ der Host Location, den Evidenztyp oder die URL des Fundorts differenzierte Informationen bereitstellen, die eine eigenständige Klassifizierung auf Ebene der Fundorte ermöglichen, welche Überschneidungen verschiedener OA-Typen für eine Publikation einbeziehen kann. Auch die eingesetzte Typologisierung sollte daher in Analysen thematisiert und bei der Interpretation von Ergebnissen berücksichtigt werden. Nicht in Vergessenheit geraten sollte schließlich, dass Unpaywall - wie im Kapitel Herangehensweise erwähnt - auf der DOI Registrierungsagentur Crossref beruht und daher nur Publikationen enthält, die bei Crossref registriert wurden. Publikationen, die keine DOI haben oder eine DOI bei einer anderen Agentur registriert haben, werden nicht aufgenommen. Einige Metadaten, wie der Dokumententyp und das Publikationsdatum werden von Crossref übernommen, wo sie wiederum von den Verlagen hinterlegt werden. Dies führt zu Inkonsistenzen, da entsprechenden Bezeichnungen nicht von allen Verlagen einheitlich verwendet werden.

Über diese Problemfelder hinaus erweisen sich die folgenden Punkte als Desiderate sowohl für die praktische Anwendung als auch im Hinblick auf wissenschaftliche Studien in der Bibliometrie. In Bezug auf verlagsbasiertes OA fehlt es momentan noch an einer Ausdifferenzierung des Modells Delayed oder Moving Wall OA für Artikel, die nicht in reinen OA-Zeitschriften veröffentlicht wurden (nach Feld *journal_is_oa*). Bei diesem Modell werden Publikationen nach Ablauf einer gewissen Zeit (Embargofrist) nach dem Publikationsdatum frei zur Verfügung gestellt. Artikel, die als Delayed OA verfügbar sind, werden zum aktuellen Zeitpunkt von Unpaywall entweder als *hybrid* (falls eine OA-Lizenz identifiziert werden konnte) oder als *bronze* (ohne erkennbare Lizenz) eingestuft. Auch hinsichtlich freier Verfügbarkeit von Publikationen in Repositorien wäre eine genaueren Unterteilung

hilfreich. Anhand von Repositorienlisten, die Repositorien, welche bestimmte Qualitäts- und Standardisierungskriterien erfüllen, gemeinsam mit Meta-Informationen über die registrierten Repositorien aufführen, könnte eine Differenzierung analog zum Feld *journal_is_in_doaj* bei verlagsbasiertem OA oder eine Spezifizierung des Repositorientyps erfolgen, beispielsweise eine Variable, die angibt ob ein Repositorium in einem entsprechenden Verzeichnis registriert ist oder den dortigen Repositorientyp auflistet. Momentan ist hierfür noch die manuelle Zusammenführung mit weiteren Nachweisquellen, wie etwa OpenDOAR notwendig. Dieses Matching gestaltet sich oft schwierig aufgrund von unterschiedlichen Datenbasen und fehlender Standardisierung der verschiedenen Quellen. Es wäre wünschenswert, wenn zukünftig Bibliotheken verstärkt an einer Erweiterung der OA-Nachweisdatenbanken arbeiten, etwa indem sie weitere Dienste integrieren oder Algorithmen für alternative Klassifikationen bereitstellen.

6 Fazit

Unpaywall ist ein breitgenutzter Datenservice mit umfangreichen Metadaten über OA-Volltexte von Crossref-indexierten Publikationen. Dieser findet sowohl im bibliothekarischen Kontext beim Publikationsmonitoring Anwendung als auch in der bibliometrischen Forschung, um aktuelle Fragestellungen bezüglich der Entwicklung des OA-Publizierens datengestützt zu beantworten. Die freie Verfügbarkeit der Dumps des Datenservices ermöglicht transparente und nachvollziehbare Datenanalysen. Mittels Google BigQuery konnten wir die einzelnen Snapshots in eine leistungsfähige Dateninfrastruktur überführen, die uns sowohl die kritische Exploration der Datenquelle als auch darauf aufbauende Analysen erlaubt. Besonderes Augenmerk liegt dabei auf der Typologisierung der OA-Nachweise, die einer fortlaufenden Änderung unterliegt. So ist auch die vorliegende Untersuchung mit der Erscheinung des aktuellsten Snapshots von Oktober 2020 mittlerweile veraltet, in dem eine neue Methode für die Differenzierung zwischen sofortigem und verzögertem OA eingeführt wurde. Diese dynamische Entwicklung Unpaywalls muss gemeinsam mit der Priorisierung des verlagsseitigen OA bei der Interpretation von Datenanalysen auf Basis von Unpaywall beachtet werden.

Literaturverzeichnis

Aspesi, Claudio; Brand, Amy (2020): In Pursuit of Open Science, Open Access Is Not Enough. In: *Science* 368 (6491): 574–577. DOI:10.1126/science.aba3763.

Barbers, Irene; Rosenberger, Sonja; Mittermaier, Bernhard (2020): Auf dem Weg zur Open Access Transformation. *Informationspraxis* 6 (2). DOI:10.11588/IP.2020.2.73240.

Else, Holly (2018): How Unpaywall Is Transforming Open Science. *Nature* 560 (7718): 290–291. DOI:10.1038/d41586-018-05968-3.

Fournier, Johannes; Weihberg, Roland (2013): Das Förderprogramm »Open Access Publizieren« der Deutschen Forschungsgemeinschaft. Zum Aufbau von Publikationsfonds an wissenschaftlichen Hochschulen in Deutschland. *Zeitschrift für Bibliothekswesen und Bibliographie* 60 (5): 236–243. DOI:10.3196/186429501360528.

Haupka, Nick; Morrison, Paul (2020): unpywall/unpywall: v0.1.9 (Version v0.1.9). Zenodo. http://doi.org/10.5281/zenodo.4085415.

Hendricks, Ginny; Tkaczyk, Dominika; Lin, Jennifer; Feeney, Patricia (2020): Crossref: The Sustainable Source of Community-Owned Scholarly Metadata. In: *Quantitative Science Studies* 1 (1): 414–27. DOI:10.1162/qss_a_00022.

Hillenkötter, Kristine; (2018): An der Schwelle zur Transformation: „alte“ und „neue“ Lizenzmodelle im Überblick. *Bibliothek Forschung und Praxis* 42 (1): 42–56. DOI:10.1515/bfp-2018-0008.

Hobert, Anne; Jahn, Najko; Mayr, Philipp; Schmidt, Birgit; Taubert, Niels (2020): Open Access Uptake in Germany 2010-18: Adoption in a diverse research landscape. Preprint verfügbar unter: DOI: 10.5281/zenodo.3892950, veröffentlicht am 15. Juni 2020.

Huang, Chun-Kai (Karl); Neylon, Cameron; Brookes-Kenworthy, Chloe; Hosking, Richard; Montgomery, Lucy; Wilson, Katie; Ozaygen, Alkim (2020): Comparison of Bibliographic Data Sources: Implications for the Robustness of University Rankings. In: *Quantitative Science Studies,* 1 (2), 445–78. DOI:10.1162/qss_a_00031.

Jahn, Najko (2019). roadoi: Find Free Versions of Scholarly Publications via Unpaywall. R package version 0.6. https://CRAN.R-project.org/package=roadoi.

Kindling, Maxi; Hampl, Martin; Finke, Pamela; Voigt, Michaela; Hübner, Andreas (2020): *Open-Access-Anteil bei Zeitschriftenartikeln von Wissenschaftlerinnen und Wissenschaftlern an Einrichtungen des Landes Berlin : Datenauswertung für das Jahr 2018.* DOI: 10.14279/depositonce-9606.

Mittermaier, Bernhard; Barbers, Irene; Ecker, Dirk; Lindstrot, Barbara; Schmiedicke, Heidi; Pollack, Philipp (2018): Der Open Access Monitor Deutschland. In: *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, 84–100. DOI:10.5282/O-BIB/2018H4S84-100.

Momeni, Fakhri; Fraser, Nicholas; Peters, Isabella; Mayr, Phillip (2019): *From closed to open access: A case study of flipped journals*. arXiv:1903.11682 [cs]. Preprint. Verfügbar unter http://arxiv.org/abs/1903.11682, veröffentlicht am 9. Oktober 2019.

Piwowar, Heather; Priem, Jason; Larivière, Vincent; Alperin, Juan Pablo; Matthias, Lisa; Norlander, Bree; Farley, Ashley; West, Jevin; Haustein, Stefanie (2018): The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles. In: *PeerJ* 6 (February). Verfügbar unter https://doi.org/10.7717/peerj.4375.

Piwowar, Heather; Priem, Jason; Orr, Richard (2019): The Future of OA: A Large-Scale Analysis Projecting Open Access Publication and Readership. Preprint. Verfügbar unter https://doi.org/10.1101/795310, veröffentlicht am 09. Oktober 2019.

Robinson-Garcia, Nicolas; Costas, Rodrigo; van Leeuwen,Thed; (2020): Open Access Uptake by Universities Worldwide. In: *PeerJ* 8: e9410. Verfügbar unter https://doi.org/10.7717/peerj.9410.

Schmeja, Stefan; Tullney, Marco (2020): Publikationsmonitoring. In: *Publikationsberatung an Universitäten*, herausgegeben von Lackner, Karin; Schilhan, Lisa; Kaier, Christian. 203–216. transcript-Verlag. DOI:10.14361/9783839450727-011.

Waltman, Ludo; Larivière, Vincent (2020): Special Issue on Bibliographic Data Sources. In: *Quantitative Science Studies* 1 (1): 360–362. DOI:10.1162/qss_e_00026.

Xia, Feng; Wang, Wei; Bekele, Teshome Megersa; Liu,Huan (2017): Big Scholarly Data: A Survey. In: *IEEE Transactions on Big Data* 3 (1): 18–35. DOI:10.1109/TBDATA.2016.2641460.

Anne Hobert
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Universität Göttingen
Platz der Göttinger Sieben 1
D-37073 Göttingen
hobert@sub.uni-goettingen.de

Nick Haupka
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Universität Göttingen
Platz der Göttinger Sieben 1
D-37073 Göttingen
nick.haupka@sub.uni-goettingen.de

Najko Jahn
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Universität Göttingen
Platz der Göttinger Sieben 1
D-37073 Göttingen
jahn@sub.uni-goettingen.de